

# Weekly Report(2018.12.17-2018.12.23)

## DONE

1. 看TPFlow论文和中文博客：
  - 当前任务
    1. 看TPFlow论文；
    2. 张量分解具体实现，其如何和深度学习结合；
    3. 调研数据挖掘中能用于解决模式识别、模式提取的算法；
    4. 学习graph embedding；
  - 日报在附录中

## 小结

- 本周开始TPFlow的算法研究，有很多概念需要学习。本校下达了毕业设计和提高实习答辩的通知，要做准备。这周周末考研结束了，今后要常常来实验室，提高自己的工作效率。

## 学习记录

学习日期	学习事项	学习时间
周一	看TPFlow	5h
周二	玉泉上课	3h
周三	看TPFlow	4h
周四	看张量	4h
周五	听组会	4h
周六	划水	0h
周日	划水	0h

## PLAN

### 短期计划

1. 继续看TPFlow。

### 中期计划

1. 自学less、webpack、promise、\$.ajax、VS Code + Git等等。
2. 能阅读购买的ES6、JavaScript标准等书的一些部分。

## 长期计划

1. 推进VIS 2019的项目进度。
2. 使用所学的网页工具实现自己的idea。

## APPENDIX

# 日报 2018.12.17

## 模式识别

这个概念之前不了解，今天整理一下。

### 基本概念

模式识别（Pattern Recognition）参考[博客-模式识别课程笔记](#)，其本质就是从大量事物中抽象整体特征，即数据的分类，有以下两种：

- 监督模式识别：已知类别、有训练集、验证集（eg. 神经网络、决策树）
- 非监督模式识别：不知道类别、根据不同的特征提取能聚类出不同的结果（eg. K-means DBScan）

### 模型方法

1. 数据获取、感知 sensing  
测量物理变量（工程中还要考虑样本质量，时间、成本因素）
2. 预处理  
数据清洗、去噪，改善数据
3. 特征提取 select features
  - [博客-图像特征提取-PCA](#)
  - [机器学习系列：（三）特征提取与处理](#)
4. 训练分类器 train classifier  
选择合适的分类算法，学习特征-模式类别的映射关系

施工中.....

## 数据挖掘、分析

翻出我以前买的书《[大数据分析:数据挖掘必备算法示例详解](#)》，回顾一下数据挖掘和数据分析的概念：

## 数据概念

### • 数据类型

不同算法适用于不同类型的数据，数据类型可以按一定规则自行转换。

- 数值型数据（连续型、离散型）：speed、hour、year
- 分类型数据（SVM二分类）：male/female
- 顺序型数据：rank0、rank1、rank2

### • 数据表示

- 每一行是一个实例case，包括所有属性在该case上的value
- 每一列是一个属性var
- 特定的一列可表示实例的标签/类别class（用于分类和预测）

## 数据分析

### • 数据分类与预测

当每个实例都有一个类别时，可以建立分类模型用于自动预测新的实例的类别，其属于**监督学习**（Supervised Learning），依类别数据不同分为两种。

- 分类问题（classification）：面向分类型数据
- 回归问题（regression）：面向数值型数据

### • 聚类分析（clustering）：对没有类别的数据按照特征进行分组

### • 推荐系统：基于user或item协同过滤，推荐符合用户特征的商品服务

### • 数据可视化：可视地呈现分析的结果用于和用户交互

## 分类模型

### • 分类模型训练

将原始数据随机分为训练集Training、验证集Validation，每次迭代训练后检测准确率Performance，然后调参。此处数据可以反复使用。

### • 分类模型使用

将测试集Test输入模型进行实验验证。此处数据只能使用一次。

## 常见分类算法

- KNN K近邻分类器
- SVM 支持向量机
- Logistic Regression 逻辑回归
- ID3、C4.5、CART 决策树算法(使用信息熵)
  - Random Forests 随机森林算法
  - GBDT 梯度提升决策树
- AdaBoost
- Naive Bayes 朴素贝叶斯分类器（概率论）

- ELM 极限学习机器（神经网络）
  - SRC 稀疏表示分类
- 可参考[博客-数据挖掘领域十大经典算法](#)

## 相关的项目经历

- 本科导师Jian-Ping Mei研究方向是数据挖掘与数据分析，让我们几个同学从文本分析、推荐系统、聚类分析等方向里选感兴趣的，然后我选了**聚类分析和数据可视化**。大二时科研项目：处理文献数据挖掘交叉领域，使用C语言文件流匹配字符串，生成关系矩阵后将节点集、边集导入**gephi**使用[社区探测算法](#)聚类，使用力导向图布局。

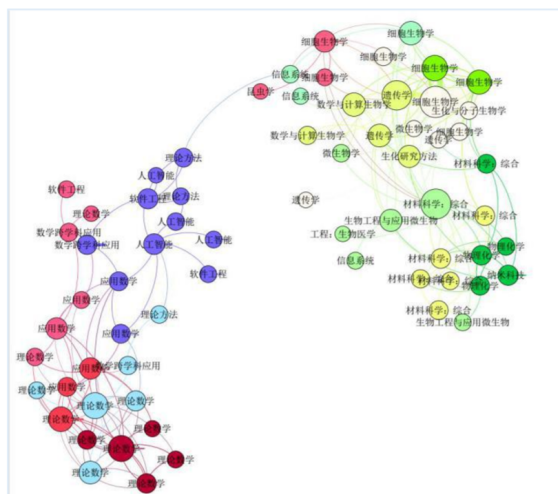


Fig. 15. Gephi Report Picture2 for journal.

### 期刊聚类

- 大三做了一年的数学建模比赛，有用到[PCA主成分分析](#)、[LASSO回归](#)、AHP层次分析法、[ARMA自回归滑动平均模型](#)等一大堆关于某种数据序列的分类、预测/取主要指标、权重的算法。

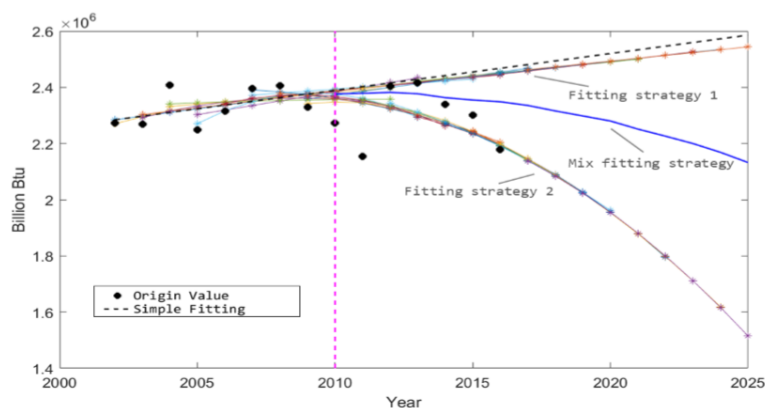


Figure 15: the fitting prediction model

### 预测模型

- 课余用Python，直接调用module做**SVM**的线性二分器、非线性二分器；自己手动实现的**SOM自组织映射神经网络**的无监督聚类二分器[百度百科-SOM](#)。（PS：TPFlow中有SOM，还没看到那里）

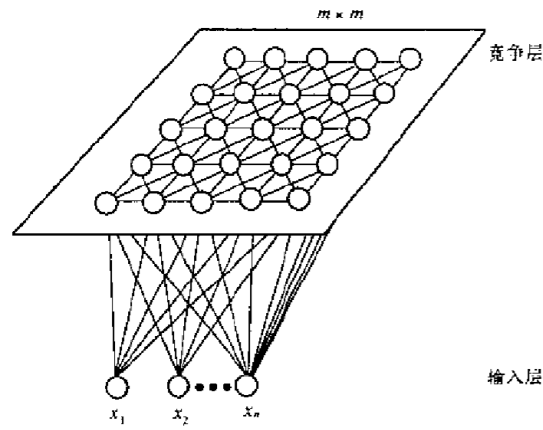


图 4.1 SOM的基本构造

#### SOM

- 保研之前做了一段时间的量化金融，用Python实现自动交易策略后**回测**（将200x-2018年6月之间的真实A股数据集导入，模拟运行），相当于**验证集Validation**，然后正式比赛实时交易就是**测试集Test**。